Varianzkomponentenschätzung

Peter von Rohr

2016-11-25

Einleitung

- BLUP-Zuchtwertschätzung: bekannte Varianzkomponenten σ_e^2 und σ_a^2
- In Praxis: Schätzung der Varianzkomponenten von Daten, Verwendung von Schätzwerten in MME
- Abgesehen von BLUP, wo sind Varianzkomponenten schon aufgetaucht?

Regression und Least Squares

- Im klassischen Regressionsmodell mit nur fixen Effekten

$$y = Xb + e \tag{1}$$

- Schätzung $\hat{b}$ mit Least Squares (kleinste Quadrate)
- Least Squares gibt keine Schätzung für σ_e^2
- Woher kommt $\widehat{\sigma_e^2}$?

Residuen

- Definition
$$r_i = y_i - \hat{y}_i = y_i - x_i^T \hat{b}$$
- Summe der quadrierten Residuen als plausibler Schätzer
$$\hat{\sigma_e}^2 = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2 \tag{2}$$
- Woher kommt Faktor $\frac{1}{n-p}$
- Grund: Erwartungstreue:
$$E\left[\hat{\sigma_e}^2\right] = \sigma_e^2$$
- Wird auch als `Least Squares` Schätzer bezeichnet

Beispiel in R

Kalb	Geschlecht	WWG
4	M	4.5
5	F	2.9
6	F	3.9
7	M	3.5
8	M	5.0

Regression in R

- Mit `summary()` werden die Resultate von `lm()` zusammengefasst

```
lmWwg <- lm(WWG ~ -1 + Geschlecht, data = dfWwgRed)
summary(lmWwg)
```

- Schätzung für $\hat{\sigma}_e$ unter `Residual standard error`

```
n <- nrow(dfWwgRed)
p <- length(unique(dfWwgRed$Geschlecht))
vecResiduals <- residuals(lmWwg)
nResVarEst <- crossprod(vecResiduals) / (n-p)
(nResSd <- sqrt(nResVarEst))
```

```
##          [,1]
## [1,] 0.745356
```

Varianzanalyse

- Ursprünglichstes Verfahren zur Schätzung von Varianzkomponenten
- Wird auch zum Testen von globalen Hypothesen verwendet
- Annahme: lineares Modell mit nur fixem Effekt b
- Frage: Haben Effektstufen von b überhaupt einen Einfluss auf y?
- Globale Nullhypothese $H_0 : b_1 = b_2 = \ldots = 0$

Tabelle der Varianzanalyse

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Geschlecht   2  79.45   39.73   71.51 0.00294 **
## Residuals    3   1.67    0.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

- Gemittelte Summenquadrate

$$MSQ_R = SSQ_R/df_R = \left(\sum_{i=1}^{n} r_i^2\right)/df_R$$

$$MSQ_b = SSQ_b/df_b = \left(\sum_{i=1}^{n} \hat{y}_i^2\right)/df_b = \left(\sum_{i=1}^{n} (x_i{}^T b)^2\right)/df_b$$

- Teststatistik mit $\mathcal{F}$-Verteilung

$$F = MSQ_b/MSQ_R$$

Schätzung von Varianzkomponenten

- "zufälliges" Modell: Beispiel Vater-Effekte

$$y = 1\mu + Zu + e$$

- Eigenschaften von u:
  - nur ein allgemeines Mittel als fixe Faktorstufe
  - Zufallsvariable mit vorgegebener Verteilung
  - unabhängig mit konstanter Varianz σ_u^2 und Erwartungswert $E[u] = 0$
- Wie können wir σ_u^2 aus den Daten schätzen?

Erwartungswerte von Summenquadraten

- Erwartungswerte von gemittelten Summenquadraten sind Funktionen von Varianzkomponenten

$$E\left[MSQ_e\right] = \sigma_e^2$$

$$E\left[MSQ_u\right] = n\sigma_u^2 + \sigma_e^2$$

- Anstelle der Erwartungswerte die empirischen Werte einsetzen

$\rightarrow$ Schätzung für Varianzkomponenten

Schätzungen

$$\widehat{\sigma_e^2} = MSQ_e$$

$$\widehat{\sigma_u^2} = \frac{MSQ_u - MSQ_e}{n}$$

Beispiel

Kalb	Vater	WWG
4	1	4.5
5	3	2.9
6	1	3.9
7	3	3.5
8	3	5.0

Tabelle der Varianzanalyse

```r
aovWwgSire <- aov(formula = WWG ~ Vater,
                  data = dfWwgSire)
summary(aovWwgSire)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Vater        1  0.192   0.192   0.229  0.665
## Residuals    3  2.520   0.840
```

Resultate

$$\widehat{\sigma_e^2} = MSQ_e = 0.84$$

Setzen wir diese Schätzung in Gleichung (**??**) ein, dann erhalten wir

$$\widehat{\sigma_u^2} = \frac{MSQ_u - MSQ_e}{n} = -0.344$$

Negative Schätzwerte

- Schätzwert für σ_u^2 ist negativ
- Ursache: spezielle Datenkonstellation (hier zu kleine Datenmenge)
- Varianzanalyse kann keine negativen Schätzwerte verhindern

$\rightarrow$ Varianzanalyse in Tierzucht kaum verwendet $\rightarrow$ andere Methoden im folgenden Kapitel